|  |  |
| --- | --- |
| **bỘ GIÁO DỤC VÀ ĐÀO TẠO** | **BỘ NÔNG NGHIỆP VÀ PTNT** |

TRƯỜNG ĐẠI HỌC THỦY LỢI



họ và tên

PHẠM MINH TIẾN

TRẦN HUY NAM

ĐỖ VĂN PHÚ

NGUYỄN BẢO QUỐC

BÁO CÁO BÀI TẬP LỚN

HỌC PHẦN PHÂN TÍCH DỮ LIỆU LỚN

ứng dụng google bigquery vào dữ đoán bệnh tiểu đường ở nữ giới

|  |  |
| --- | --- |
| NGƯỜI HƯỚNG DẪN | TS. Tạ Quang Chiếu |

Hà Nội, năm 2023

MỤC LỤC

[DANH MỤC CÁC HÌNH ẢNH iv](#_Toc134142162)

[DANH MỤC BẢNG BIỂU v](#_Toc134142163)

[DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ vi](#_Toc134142164)

[CHƯƠNG 1 GIỚI THIỆU BÀI TOÁN 1](#_Toc134142165)

[CHƯƠNG 2 CỞ SỞ LÝ THUYẾT 2](#_Toc134142166)

[2.1 Google BigQuery 2](#_Toc134142167)

[2.1.1 Google BigQuery là gì 2](#_Toc134142168)

[2.1.2 Cấu trúc của BigQuery 2](#_Toc134142169)

[2.1.3 Phương Thức Vận Chuyển Data ETL Và ELT 4](#_Toc134142170)

[2.1.4 Đặc trưng của BigQuery 6](#_Toc134142171)

[2.1.5 Pricing 8](#_Toc134142172)

[2.1.6 Phân Tích Mô Hình Định Giá 8](#_Toc134142173)

[2.1.7 Định Giá Lưu Trữ 8](#_Toc134142174)

[2.1.8 Tài Khoản Dịch Vụ (Service Account) 8](#_Toc134142175)

[2.1.9 Ưu điểm và nhược điểm của BigQuery 9](#_Toc134142176)

[2.1.10 Những lợi ích mà BigQuery đem lại: 10](#_Toc134142177)

[2.2 Hồi quy Logistic 10](#_Toc134142178)

[2.2.1 Hồi quy Logistic là gì ? 10](#_Toc134142179)

[2.2.2 Công thức hồi quy Logistic 10](#_Toc134142180)

[2.2.3 Các loại hồi quy Logistic 11](#_Toc134142181)

[2.2.4 Các thành phần của hồi quy Logistic 12](#_Toc134142182)

[2.2.5 Cách hoạt động của hồi quy Logistic 12](#_Toc134142183)

[2.2.6 Mục đích của hồi quy Logistic 13](#_Toc134142184)

[2.2.7 Sự khác biệt của hồi quy Logistic với các kỹ thuật Machine Learning khác 13](#_Toc134142185)

[2.2.8 Một số ưu điểm và nhược điểm của việc sử dụng hồi quy logistic so với các kỹ thuật Machine Learning khác 14](#_Toc134142186)

[2.2.9 Ứng dụng của hồi quy Logistic 15](#_Toc134142187)

[CHƯƠNG 3 Kết quả và đánh giá mô hình 17](#_Toc134142188)

[3.1 Phân tích chi tiết bài toán 17](#_Toc134142189)

[3.1.1 Tập dữ liệu 17](#_Toc134142190)

[3.2 Các công cụ, thư viện dùng xây dựng 17](#_Toc134142191)

[3.3 Xây dựng mô hình sử dụng BigQuery UI 19](#_Toc134142192)

[3.3.1 Đưa dữ liệu lên Google BigQuery 19](#_Toc134142193)

[3.3.2 Xây dựng mô hình 21](#_Toc134142194)

[3.3.3 Chạy thử mô hình 22](#_Toc134142195)

[3.3.4 Đánh giá độ chính xác của mô hình 23](#_Toc134142196)

[3.3.5 Ma trận nhầm lẫn 24](#_Toc134142197)

[Kết Luận 25](#_Toc134142198)

[TÀI LIỆU THAM KHẢO 26](#_Toc134142199)

[PHỤ LỤC 27](#_Toc134142200)

DANH MỤC CÁC HÌNH ẢNH

[Hình 2.1 Dữ liệu đưa vào 2](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142148)

[Hình 2.2 Các công nghệ cơ sở hạ tầng cấp thấp nhất của Google 3](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142149)

[Hình 2.3 Phương thức vận chuyển đa ta ETL 5](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142150)

[Hình 2.4 Phương thức vận chuyển ELT 5](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142151)

[Hình 3.1 Chọn ADD 19](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142152)

[Hình 3.2 Chọn Local file 19](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142153)

[Hình 3.3 Điền thông tin table 20](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142154)

[Hình 3.4 Sau khi đưa dữ liệu lên thành công 20](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142155)

[Hình 3.5 Tổng quan về model 21](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142156)

[Hình 3.6 Tổng quan về model 21](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142157)

[Hình 3.7 Tổng quan về model 22](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142158)

[Hình 3.8 Kết quả thực thi 23](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142159)

[Hình 3.9 Kết quả thực thi 24](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142160)

[Hình 3.10 Kết quả thực thi 24](https://wru-my.sharepoint.com/personal/1951061055_e_tlu_edu_vn/Documents/Huong%20dan%20trinh%20bay%20báo%20cáo%20project.docx#_Toc134142161)

DANH MỤC BẢNG BIỂU

DANH MỤC CÁC TỪ VIẾT TẮT VÀ GIẢI THÍCH CÁC THUẬT NGỮ

**CSV** Comma-Separated Values

**DDL** Data Definition Language

**DML** Data Manipulation Language

**ETL** Extract, Transform, Load

**ELT** Extract, Load, Transform

**JDBC** Java Database Connectivity

**JSON** JavaScript Object Notation

**ML** Machine Learning

**ODBC** Open Database Connectivity

**ORC** Optimized Row Columnar

**RDBMS** Relational Database Management System

**SQL** Structured Query Language

**UI** User Interface

# GIỚI THIỆU BÀI TOÁN

Google BigQuery là một dịch vụ lưu trữ và xử lý dữ liệu lớn trên đám mây của Google. Nó được thiết kế để hỗ trợ việc truy vấn, phân tích và trích xuất thông tin từ các tập dữ liệu lớn bằng cách sử dụng SQL và công nghệ xử lý phân tán.

BigQuery được tích hợp sẵn với các công cụ thống kê và khai phá dữ liệu của Google, bao gồm Google Analytics, Google Sheets, Google AdWords và Google Cloud Storage. Nó cũng có thể được tích hợp với các công cụ khác như Tableau, QlikView và Excel thông qua kết nối JDBC/ODBC.

Các tính năng chính của BigQuery bao gồm:

* Khả năng lưu trữ và xử lý dữ liệu lớn
* Khả năng tìm kiếm, truy vấn và phân tích dữ liệu bằng SQL
* Sự phân tán và tự động mở rộng để xử lý các tập dữ liệu lớn

BigQuery được sử dụng rộng rãi trong các lĩnh vực như phân tích dữ liệu, khai thác dữ liệu, kho dữ liệu và phát triển ứng dụng.

Với đề tài dự đoán mắc bệnh tiểu đường ở nữ giới, việc áp dụng BigQuery mang lại nhiều lợi ích bao gồm phát hiện sớm các trường hợp mắc bệnh tiểu đường ở nữ giới và giúp các chuyên gia y tế quản lý và điều trị bệnh hiệu quả hơn.

# CỞ SỞ LÝ THUYẾT

## Google BigQuery

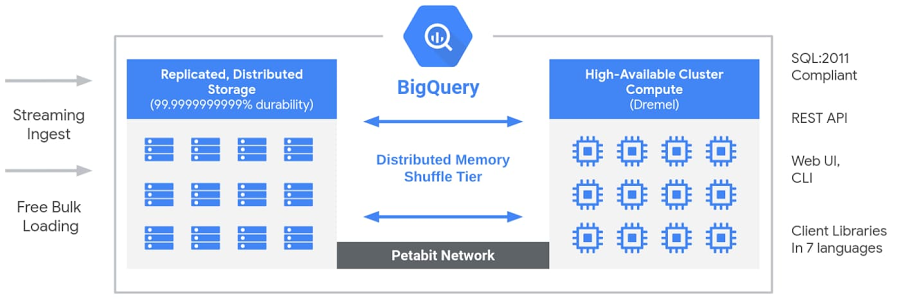
### Google BigQuery là gì

Google BigQuery là một kho dữ liệu (data warehouse) được phát hành và phổ biến vào năm 2011 trên nền tảng điện toán đám mây của Google (Google Cloud) cho phép bạn chạy các truy vấn siêu nhanh trên các tập dữ liệu lớn. BigQuery hoạt động dưới dạng nền tảng là một dịch vụ (Platform as a Service – PaaS) trên hạ tầng của Google nên được thừa hưởng nhiều công nghệ hiện đại và tiên tiến. BigQuery hiện nay đã trở thành một trong những data warehouse được sử dụng nhiều nhất trên thế giới.

SQL kết hợp với cấu trúc hạ tầng mạnh mẽ của Google giúp người dùng quản lý, xem và thực hiện các truy xuất dữ liệu từ nhiều nguồn khác nhau.

Bạn có thể xuất dữ liệu phiên hoạt động và dữ liệu lượt truy cập từ tài khoản Google Analytics 360 sang BigQuery, sau đó sử dụng cú pháp giống như SQL để truy vấn tất cả dữ liệu trong Analytics. Sử dụng Machine Learning có thể giúp các Data Scientist và Data Analyst xây dựng và sử dụng được nhiều module bằng những cú pháp SQL đơn giản. Ngoài ra, BigQuery được sử dụng rộng rãi bởi các công ty: UPS, Twitter, và Dow Jones…

### Cấu trúc của BigQuery

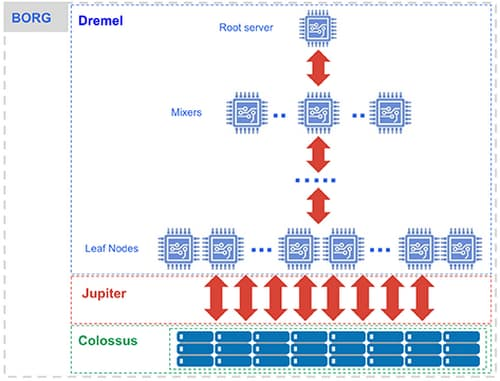


Hình . Dữ liệu đưa vào

Dữ liệu đưa vào được chia làm 2 loại: dữ liệu được đưa vào liên tục trong một luồng dữ liệu (streaming ingest) và dữ liệu được đưa vào theo một khối (free bulk loading). Hai loại dữ liệu này tách biệt với nhau nên có thể scale up chúng một cách độc lập tùy theo nhu cầu sử dụng và giúp kiểm soát phần coding rõ ràng hơn.

Các loại định dạng dữ liệu đầu vào mà BigQuery có thể đọc được: JSON (đã nén), CSV (đã nén), JSON, CSV, Parquet/ORC, Avro (chưa nén), Avro (đã nén).

Bên cạnh đó, BigQuery sử dụng một loạt các dịch vụ đa nhiệm được thúc đẩy bởi các công nghệ cơ sở hạ tầng cấp thấp nhất của Google như Dremel, Colossus, Jupiter và Bord.



Hình . Các công nghệ cơ sở hạ tầng cấp thấp nhất của Google

* **Tính toán bằng**[**Dremel**](https://research.google/pubs/pub36632/)**– một cụm đa nhiệm lớn thực thi các truy vấn SQL**

Dremel biến các truy vấn SQL thành cây thực thi. Các lá của cây được gọi là slot và thực hiện nhiệm vụ đọc dữ liệu từ kho lưu trữ và những tính toán cần thiết. Các nhánh cây được gọi là mixer, là nơi tập hợp các dữ liệu lại với nhau.

Dremel tự động sắp xếp vị trí cho các truy vấn tùy theo mức độ cần thiết, duy trì mức độ ưu tiên tương đương cho các truy vấn đồng thời từ nhiều người dùng. Một người dùng có thể nhận được hàng nghìn vị trí để chạy các truy vấn của họ.

* **Bộ nhớ là**[**Colossus**](https://www.systutorials.com/3202/colossus-successor-to-google-file-system-gfs/) **– hệ thống lưu trữ toàn cầu của Google**

BigQuery tận dụng định dạng lưu trữ theo cột và thuật toán nén để lưu trữ dữ liệu trong Colossus, được tối ưu hóa để đọc một lượng lớn dữ liệu có cấu trúc.

Colossus cũng xử lý sao chép, phục hồi (khi đĩa gặp sự cố) và quản lý phân tán (do đó sẽ không có điểm lỗi nào). Colossus cho phép người dùng BigQuery mở rộng quy mô lên hàng chục petabyte dữ liệu được lưu trữ liền mạch mà không phải trả tiền phạt khi đính kèm tài nguyên máy tính đắt hơn nhiều so với kho dữ liệu truyền thống.

* **Máy tính và bộ nhớ giao tiếp với nhau thông qua mạng petabit**[Jupiter](https://cloudplatform.googleblog.com/2015/06/A-Look-Inside-Googles-Data-Center-Networks.html)

Giữa bộ nhớ và bộ xử lý tính toán có một shuffle với nhiệm vụ sử dụng mạng Jupiter của Google để di chuyển dữ liệu từ nơi này sang nơi khác với tốc độ cực kỳ nhanh chóng.

* **BigQuery được điều phối thông qua**[**Bord**](https://research.google/pubs/pub43438/)**– tiền thân của**[**Kubernetes**](https://kubernetes.io/)

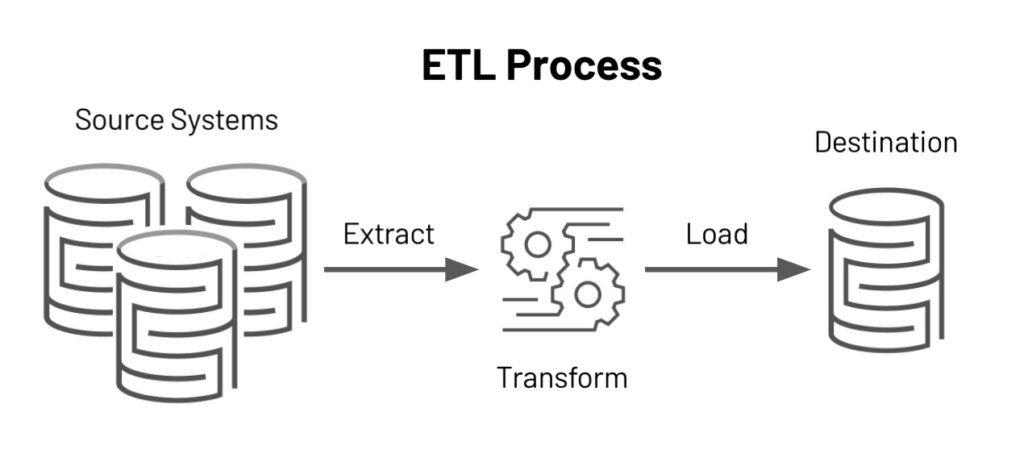
Borg phân bổ tài nguyên phần cứng và điều hành các mixer và slot.

Google rất chú trọng vào việc liên tục cải tiến những công nghệ này. Người dùng BigQuery sẽ được hưởng lợi từ những cải tiến liên tục về hiệu suất, độ bền, hiệu quả và khả năng mở rộng mà không cần ngừng hoạt động để nâng cấp như theo công nghệ truyền thống.

### Phương Thức Vận Chuyển Data ETL Và ELT

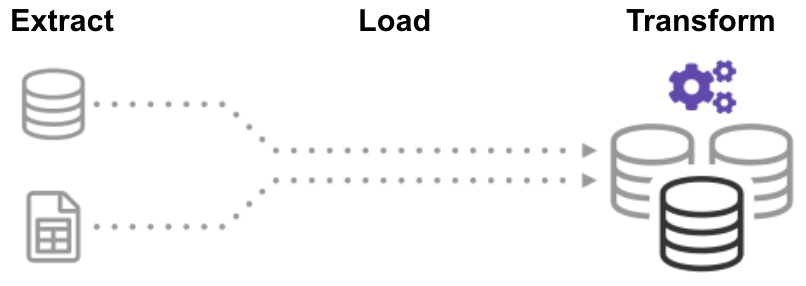
ETL (Extract, Transform, Load) và ELT (Extract, Load, Transform) đều là các phương pháp tích hợp dữ liệu giúp chuyển dữ liệu từ một nguồn bên ngoài vào data warehouse.

ETL là tên viết tắt của Extract, Transform và Load. Trong hệ thống Data Warehouse, một công cụ ETL trích xuất dữ liệu từ các hệ thống nguồn RDBMS khác nhau sau đó chuyển đổi dữ liệu như áp dụng các biến đổi dữ liệu ( tính toán, nối chuỗi v.v. ) và sau đó tải dữ liệu vào hệ thống Data Warehouse. ETL là những luồng từ “nguồn” tới ”đích”. Trong quá trình ETL, engine chuyển đổi sẽ xử lý mọi thay đổi dữ liệu. Quá trình ETL bắt buộc cần quá trình “Transform” hoàn tất trước ki dữ liệu được lưu (Load) vào kho. Khi kích thước dữ liệu tăng lên, thời gian chuyển đổi cũng tăng theo.



Hình . Phương thức vận chuyển đa ta ETL

ELT là một phương pháp khác để tiếp cận công cụ chuyển động dữ liệu trong data warehouse. Thay vì chuyển đổi dữ liệu trước khi lưu vào kho, ELT cho phép người sử dụng lưu trực tiếp dữ liệu thô (Raw Data) vào kho. Nhờ sức mạnh xử lý của Google mà ELT có thể sử dụng cho lượng data rất lớn mà không ảnh hưởng quá nhiều về hiệu suất.



Hình . Phương thức vận chuyển ELT

BigQuery chú trọng sử dụng ELT hơn ETL bởi vì ELT dễ sử dụng hơn, giảm bớt khối lượng công việc cho Data Engineer.

### Đặc trưng của BigQuery

**Tính linh hoạt, giá cả có thể dự đoán và hiệu suất giá tốt nhất:** [Các phiên bản BigQuery](https://cloud.google.com/bigquery/docs/editions-intro) cho phép bạn chọn bộ tính năng phù hợp cho các yêu cầu về khối lượng công việc riêng lẻ với khả năng kết hợp và kết hợp để có hiệu suất-giá phù hợp. Tính năng tự động thay đổi quy mô công suất điện toán bổ sung các tài nguyên điện toán chi tiết trong thời gian thực để phù hợp với nhu cầu khối lượng công việc của bạn và đảm bảo bạn chỉ trả tiền cho công suất điện toán mà bạn sử dụng. Với giá lưu trữ nén, bạn có thể giảm chi phí lưu trữ đồng thời tăng mức sử dụng dữ liệu của mình.

**Học máy tích hợp**: BigQuery ML cho phép các nhà khoa học dữ liệu và nhà phân tích dữ liệu xây dựng và vận hành các mô hình ML trên dữ liệu có cấu trúc, bán cấu trúc và hiện tại ở quy mô toàn cầu trực tiếp bên trong BigQuery, sử dụng SQL đơn giản—trong một khoảng thời gian ngắn. Xuất các mô hình BigQuery ML để dự đoán trực tuyến vào Vertex AI hoặc lớp phục vụ của riêng bạn

**Phân tích và chia sẻ dữ liệu trên các đám mây:** [BigQuery Omni là một giải pháp phân tích đa đám mây](https://cloud.google.com/bigquery/docs/omni-introduction) , được quản lý hoàn toàn, cho phép phân tích dữ liệu an toàn và tiết kiệm chi phí trên các đám mây và chia sẻ kết quả trong một ô kính duy nhất. Trong Trung tâm phân tích BigQuery, [trao đổi](https://cloud.google.com/analytics-hub) nội bộ và giữa các tổ chức một cách an toàn mà không cần di chuyển dữ liệu, đồng thời nâng cao khả năng phân tích với [bộ dữ liệu thương mại, công khai và Google](https://cloud.google.com/datasets) .

**Phân tích thời gian thực với các đường ống truyền dữ liệu:** BigQuery có các khả năng tích hợp để nhập dữ liệu truyền trực tuyến và cung cấp dữ liệu đó ngay lập tức để truy vấn, cùng với các tích hợp riêng cho các sản phẩm truyền trực tuyến như [Dataflow](https://cloud.google.com/dataflow) . Phân tích các tập dữ liệu lớn một cách tương tác với BigQuery BI Engine, một [dịch vụ phân tích trong bộ nhớ](https://cloud.google.com/bi-engine/docs) cung cấp thời gian phản hồi truy vấn dưới giây và tính đồng thời cao. Tăng tốc hiệu suất truy vấn và giảm chi phí trong môi trường của bạn với  [các chế độ xem cụ thể hóa BigQuery](https://cloud.google.com/bigquery/docs/materialized-views) .

**Thống nhất, quản lý và chi phối tất cả các loại dữ liệu:** Truy vấn tất cả các loại dữ liệu với BigQuery: có cấu trúc, bán cấu trúc và không có cấu trúc. Sử dụng BigLake để khám phá và [thống nhất các loại dữ liệu khác nhau](https://cloud.google.com/biglake) cũng như xây dựng các mô hình nâng cao. Khám phá, quản lý, giám sát và quản lý dữ liệu một cách tập trung trên [các hồ dữ liệu](https://cloud.google.com/dataproc) , kho dữ liệu và siêu thị dữ liệu bằng các biện pháp kiểm soát nhất quán với Dataplex, một [kết cấu dữ liệu thông minh](https://cloud.google.com/dataplex) cho phép các tổ chức cung cấp quyền truy cập vào dữ liệu đáng tin cậy.

**Chia sẻ thông tin chuyên sâu với thông tin kinh doanh tích hợp sẵn:** Với thông tin kinh doanh tích hợp sẵn, hãy tạo và chia sẻ thông tin chi tiết chỉ bằng vài cú nhấp chuột với [Looker Studio](https://lookerstudio.google.com/overview) hoặc xây dựng trải nghiệm giàu dữ liệu vượt xa BI với [Looker](https://cloud.google.com/looker) . Phân tích hàng tỷ hàng dữ liệu BigQuery trực tiếp trong Google Trang tính bằng các công cụ quen thuộc như bảng tổng hợp, biểu đồ và công thức để dễ dàng nhận được thông tin chi tiết từ dữ liệu lớn với [Trang tính được liên kết](https://cloud.google.com/blog/products/g-suite/connected-sheets-is-generally-available) .

**Quản trị và bảo mật dữ liệu:** Việc tích hợp BigQuery  [với các dịch vụ bảo mật và quyền riêng tư](https://cloud.google.com/bigquery/docs/access-control)  từ Google Cloud cung cấp các biện pháp kiểm soát quản trị chi tiết và bảo mật mạnh mẽ, xuống  [cấp độ cột](https://cloud.google.com/bigquery/docs/column-level-security-intro)  và  [cấp độ hàng](https://cloud.google.com/bigquery/docs/row-level-security-intro) . Hãy yên tâm khi biết rằng dữ liệu của bạn được  [mã hóa](https://cloud.google.com/bigquery/docs/encryption-at-rest)  ở trạng thái lưu trữ và chuyển tiếp theo mặc định.

**Phân tích không gian địa lý với BigQuery:** [Không gian địa lý BigQuery](https://cloud.google.com/bigquery/docs/geospatial-intro) kết hợp độc đáo kiến ​​trúc serverless của BigQuery với hỗ trợ riêng cho phân tích không gian địa lý, vì vậy bạn có thể tăng cường quy trình phân tích của mình bằng thông tin vị trí. Đơn giản hóa các phân tích của bạn, xem dữ liệu không gian theo những cách mới và mở khóa các ngành kinh doanh hoàn toàn mới với sự hỗ trợ cho các điểm, đường, đa giác và đa giác tùy ý ở các định dạng dữ liệu không gian địa lý phổ biến.

**Thu thập và sao chép dữ liệu thay đổi theo thời gian thực:** Đồng bộ hóa dữ liệu trên các cơ sở dữ liệu, hệ thống lưu trữ và ứng dụng không đồng nhất một cách đáng tin cậy và có độ trễ tối thiểu với  [Datastream](https://cloud.google.com/datastream) . [Luồng dữ liệu tích hợp với các mẫu Luồng dữ liệu](https://cloud.google.com/dataflow/docs/guides/templates/provided-streaming" \l "datastream-to-bigquery) có thể mở rộng và được xây dựng có mục đích   để kéo các luồng thay đổi được ghi vào Bộ lưu trữ đám mây và tạo các bảng sao chép cập nhật trong BigQuery để phân tích theo thời gian thực.

**SQL chuẩn:** BigQuery hỗ trợ một phương ngữ SQL tiêu chuẩn tuân thủ ANSI 2011, giúp giảm nhu cầu viết lại mã. BigQuery cũng cung cấp trình điều khiển ODBC và JDBC miễn phí để đảm bảo các ứng dụng hiện tại của bạn có thể tương tác với công cụ mạnh mẽ của nó.

### Pricing

Định giá phân tích ([Analysis pricing](https://cloud.google.com/bigquery/pricing#analysis_pricing_models)) là chi phí để xử lý các truy vấn, bao gồm truy vấn SQL, hàm do người dùng xác định, tập lệnh và một số câu lệnh ngôn ngữ thao tác dữ liệu (DML) và ngôn ngữ định nghĩa dữ liệu (DDL).

Định giá bộ nhớ ([Storage pricing](https://cloud.google.com/bigquery/pricing#storage)) là chi phí để lưu trữ dữ liệu mà bạn tải vào BigQuery.

### Phân Tích Mô Hình Định Giá

Định giá dựa trên nhu cầu [(on demand pricing)](https://cloud.google.com/bigquery/pricing#on_demand_pricing): với mô hình định giá này, phí được tính theo số lượng bytes được xử lý cho mỗi lần query. 1 TB dữ liệu query xử lý đầu tiên vào mỗi tháng sẽ được miễn phí.

Định giá thống nhất [(flat-rate pricing)](https://blog.hubspot.com/sales/flat-rate-pricing): với mô hình định giá này, khi mua slots, cũng chính là các CPUs ảo (virtual CPUs), là đang mua khả năng xử lý chuyên dụng để có thể sử dụng trong việc chạy query.

### Định Giá Lưu Trữ

Định giá lưu trữ là chi phí để lưu trữ data load vào BigQuery.

Lưu trữ ngắn hạn [(Active storage)](https://cloud.google.com/bigquery/pricing#on_demand_pricing) bao gồm table và table partition đã được chỉnh sửa trong vòng 90 ngày.

Lưu trữ lâu dài [(Long-term storage)](https://cloud.google.com/bigquery/pricing#on_demand_pricing) là table và table partition không được chỉnh sửa trong vòng 90 ngày liên tiếp.

Miễn phí 10GB đầu/ tháng.

### Tài Khoản Dịch Vụ (Service Account)

Nếu bạn muốn tích hợp BigQuery vào một công cụ thứ ba, chúng ta phải sử dụng một tài khoản Dịch vụ để hệ thống nhận diện chúng ta là key pair để kết nối. Đây là tài khoản đặc biệt được sử dụng bởi một ứng dụng hoặc phần mềm chứ không phải một người.

Tài khoản dịch vụ có thể thực hiện các lệnh gọi API được ủy quyền với tư cách là tài khoản dịch vụ hoặc là người dùng Google Workspace, Cloud Identity thông qua ủy quyền trên toàn miền.

### Ưu điểm và nhược điểm của BigQuery

* Ưu điểm:

Khả năng xử lý dữ liệu lớn: BigQuery được thiết kế để xử lý các tập dữ liệu lớn đến hàng tỷ hàng triệu hàng tỉ dòng. Nó có thể xử lý các tệp có dung lượng lên đến hàng petabyte.

Khả năng tích hợp dữ liệu dễ dàng: BigQuery có khả năng tự động kết hợp các nguồn dữ liệu khác nhau và hiển thị chúng trong cùng một bảng. Điều này giúp cho việc truy xuất và phân tích dữ liệu trở nên dễ dàng hơn.

Tốc độ xử lý nhanh: BigQuery sử dụng công nghệ "columnar storage" để lưu trữ dữ liệu, giúp cho việc truy xuất dữ liệu trở nên nhanh chóng hơn.

Hỗ trợ nhiều ngôn ngữ lập trình: BigQuery hỗ trợ nhiều ngôn ngữ lập trình như Python, Java, Ruby và Go. Điều này giúp các nhà phát triển có thể xây dựng ứng dụng và tích hợp dữ liệu với BigQuery dễ dàng hơn.

An toàn và bảo mật: BigQuery hỗ trợ các tính năng bảo mật như quản lý danh sách đen/ trắng, mã hóa dữ liệu và phân quyền người dùng.

* Nhược điểm:

Giá thành cao: BigQuery là một sản phẩm đắt tiền và tính phí dựa trên khối lượng dữ liệu được xử lý.

Không hỗ trợ các truy vấn phức tạp: BigQuery không hỗ trợ một số loại truy vấn phức tạp như truy vấn theo bộ lọc nhóm hoặc tính toán liên quan đến thời gian.

Phụ thuộc vào kết nối internet: Việc sử dụng BigQuery phụ thuộc vào kết nối Internet để truy xuất dữ liệu, do đó, nếu kết nối chậm hoặc bị gián đoạn, thời gian truy xuất dữ liệu sẽ chậm hơn.

Không hỗ trợ cập nhật dữ liệu real-time: BigQuery không hỗ trợ cập nhật dữ liệu real-time và chỉ cho phép cập nhật dữ liệu theo batch.

Không phù hợp cho các ứng dụng với tần suất truy xuất dữ liệu cao: BigQuery được thiết kế cho các truy vấn phân tích dữ liệu, và không phù hợp cho các ứng dụng với tần suất truy xuất dữ liệu cao.

### Những lợi ích mà BigQuery đem lại:

* Tự động phân phối dữ liệu
* Tăng khả năng tiếp cận các insights
* Xây dựng nền tảng cho trí tuệ nhân tạo (AI)
* Chuẩn bị các insights đúng thời điểm
* Bảo mật dữ liệu kinh doanh
* Đơn giản hóa quy trình dữ liệu hoạt động

## Hồi quy Logistic

### Hồi quy Logistic là gì ?

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

Nó được dùng phổ biến trong phân tích các dữ liệu và khai thác thông tin dữ liệu, đặc biệt là trong các lĩnh vực có liên quan đến dự đoán kết quả và phân loại số liệu.

### Công thức hồi quy Logistic

Trong hồi quy logistic, logit được tính bằng xác suất thành công chia cho xác suất thất bại. Hàm hồi quy logistic có các công thức sau:

Trong đó: ***P*** là nhân tố xác suất của biến quan sát một trường hợp ***i*** trong kết quả biến ***Y*** với một giá trị là bằng 1; ***e*** là một hằng số toán học Euler với giá trị gần bằng 2.71828 và các hệ số hồi quy tương ứng với các biến quan sát

Chúng ta thường sử dụng mô hình hồi quy Logistic để ước lượng sự ảnh hưởng của các biến X về một Odds (Y=1).

Hồi quy logistic có thể được dùng cho cả trường hợp có một biến độc lập (hồi quy logistic đơn biến) và trường hợp có nhiều biến độc lập (hồi quy logistic đa biến).

### Các loại hồi quy Logistic

Có ba loại mô hình hồi quy logistic là:

* Hồi quy logistic nhị phân

Với hồi quy logistic nhị phân (Binary logistic regression), kết quả hoặc biến phụ thuộc bản chất nhị phân - tức là chỉ có hai kết quả có thể xảy ra (ví dụ: 0 hoặc 1). Một số vai trò phổ biến của loại mô hình này gồm dự đoán email là spam hay không phải spam, khối u là ác tính hay không ác tính. Trong hồi quy logistic, hồi quy logistic nhị phân là cách tiếp cận được sử dụng nhiều nhất.

* Hồi quy logistic đa thức

Trong hồi quy logistic đa thức (Multinomial logistic regression), biến phụ thuộc có 3 hoặc nhiều kết quả có thể xảy ra; tuy nhiên, các kết quả không có thứ tự cụ thể.

Ví dụ: Các hãng phim muốn dự đoán thể loại phim mà khán giả có thể xem để marketing hiệu quả hơn. Mô hình hồi quy logistic đa thức có thể giúp hãng phim xác định mức ảnh hưởng của tuổi tác, giới tính và tình trạng quan hệ của một người đến loại phim người đó yêu thích. Sau đó, hãng có thể định hướng chiến dịch quảng cáo phim đến nhóm người có khả năng đi xem bộ phim đó hất.

* Hồi quy logistic thứ tự

Trong hồi quy logistic thứ tự (Ordinal logistic regression), biến phụ thuộc có ba hoặc nhiều kết quả có thể xảy ra, nhưng trong trường hợp này, các kết quả có thứ tự xác định. Ví dụ: thang điểm từ A đến F hoặc thang đánh giá từ 1 đến 5.

### Các thành phần của hồi quy Logistic

**Hàm sigmoid:** là một hàm số phi tuyến, được sử dụng để chuyển đổi các giá trị đầu vào thành các giá trị xác suất ở đầu ra trong khoảng từ 0 đến 1.

Biến độc lập và biến phụ thuộc

**Biến phụ thuộc:** đây là biến mà chúng ta muốn dự đoán và có giá trị rời rạc hoặc nhị phân (thường được mã hóa dưới dạng 0 hoặc 1).

**Biến độc lập:** là các yếu tố đầu vào được sử dụng để dự đoán biến phụ thuộc. Các biến đó là các biến định lượng (quantitative variables) hoặc các biến định tính (qualitative variables).

**Tham số số hồi quy:** là các giá trị được ước tính trong quá trình huấn luyện mô hình, thể hiện mức độ ảnh hưởng của mỗi biến độc lập đến biến phụ thuộc.

**Hàm logistic**: hàm này biểu diễn mối quan hệ giữa các biến độc lập và biến phụ thuộc. Nó chuyển đổi đầu vào liên tục thành một giá trị rời rạc hoặc nhị phân.

**Hàm mất mát:** hàm này đánh giá sự sai khác giữa giá trị dự đoán và giá trị thực tế của biến phụ thuộc, và được sử dụng để tối ưu hóa các tham số của mô hình.

Hệ số tương quan r là một khái niệm khá quen thuộc trong hoạt động xử lý và phân tích số liệu spss. Mỗi giá trị và mỗi biến tính hệ số tương quan r đều mang một ý nghĩa khác nhau và thể hiện chức năng riêng của nó.

### Cách hoạt động của hồi quy Logistic

**Xác định câu hỏi**: Mọi quá trình phân tích dữ liệu đều bắt đầu bằng một câu hỏi kinh doanh. Đối với hồi quy logistic, các câu hỏi nên được giới hạn trong phạm vi nhất định để kết quả được cụ thể. Chẳng hạn như những ngày mưa có ảnh hưởng đến doanh thu hàng tháng của công ty hay không? (câu trả lời có hoặc không)

**Thu thập dữ liệu lịch sử**: Sau khi xác định câu hỏi, bước tiếp theo là xác định các yếu tố dữ liệu liên quan và thu thập dữ liệu trước đây cho tất cả yếu tố. Ví dụ: Để trả lời câu hỏi “Những ngày mưa có ảnh hưởng đến doanh thu hàng tháng của công ty hay không?”, tổ chức có thể cần thu thập dữ liệu về doanh số hàng tháng cùng với số ngày mưa mỗi tháng trong ba năm qua.

**Đào tạo mô hình phân tích hồi quy**: Dữ liệu lịch sử sẽ được xử lý thông qua phần mềm hồi quy. Phần mềm hồi quy sẽ xử lý các điểm dữ liệu và kết nối chúng bằng phương trình.

**Dự đoán các giá trị không xác định:** Đối với các giá trị không xác định, phần mềm sẽ áp dụng phương trình để dự đoán

### Mục đích của hồi quy Logistic

Hồi quy logistic có thể đo lường tác động của nhiều biến số (ví dụ: tuổi, giới tính, vị trí đặt quảng cáo) với một kết quả nhất định (ví dụ: nhấp hoặc bỏ qua). Thông qua đó, giúp phân tích hiệu quả tương đối của các phương án khác nhau đối với nhóm người khác nhau, như thanh niên/già hoặc nam/nữ.

Trong học máy (Machine learning), hồi quy logistic thường được sử dụng cho các bài toán phân loại nhị phân, là các bài toán có hai lớp giá trị, gồm các dự đoán như "cái này hoặc cái kia", "có hoặc không" và "A hoặc B".

Hồi quy logistic cũng có thể ước tính xác suất của các sự kiện, như xác định mối quan hệ giữa các thông tin đặc trưng và xác suất của kết quả. Chẳng hạn, hồi quy logistic có thể tạo ra một mô hình tương quan giữa số giờ học với khả năng đạt hoặc trượt của sinh viên. Mặt khác, mô hình tương tự có thể được sử dụng để dự đoán liệu một sinh viên sẽ đạt hay không dựa trên số giờ học và kết quả là: đạt hoặc không đạt.

### Sự khác biệt của hồi quy Logistic với các kỹ thuật Machine Learning khác

Hai kỹ thuật phân tích dữ liệu phổ biến là phân tích hồi quy tuyến tính và học sâu.

* Hồi quy logistic so với hồi quy tuyến tính

Hồi quy tuyến tính dự đoán một biến phụ thuộc liên tục bằng một tập hợp các biến độc lập cho trước. Một biến liên tục có thể có một phạm vi giá trị, chẳng hạn như giá cả hoặc độ tuổi. Do đó, hồi quy tuyến tính có thể dự đoán giá trị thực của biến phụ thuộc. Kỹ thuật này có thể trả lời các câu hỏi như "Giá gạo sau 10 năm nữa sẽ là bao nhiêu?" Không giống như hồi quy tuyến tính, hồi quy logistic là một thuật toán phân loại. Kỹ thuật này không thể dự đoán giá trị thực sự cho dữ liệu liên tục. Kỹ thuật này có thể trả lời các câu hỏi như "Liệu giá gạo trong 10 năm nữa có tăng 50% hay không?"

* Hồi quy logistic so với học sâu

Hồi quy logistic ít phức tạp và có cường độ điện toán ít hơn so với học sâu. Quan trọng hơn là nhà phát triển không thể điều tra hoặc sửa đổi các phép toán học sâu do tính chất phức tạp và dựa trên máy móc của chúng. Mặt khác, các phép toán hồi quy logistic lại minh bạch và dễ khắc phục sự cố hơn.

### Một số ưu điểm và nhược điểm của việc sử dụng hồi quy logistic so với các kỹ thuật Machine Learning khác

* Ưu điểm

**Tính đơn giản**: Các mô hình hồi quy logistic ít phức tạp về mặt toán học hơn các phương pháp Machine Learning khác. Do đó, bạn có thể triển khai chúng ngay cả khi đội ngũ của bạn không ai có chuyên môn sâu về Machine Learning.

**Tốc độ**: Các mô hình hồi quy logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao bởi chúng cần ít khả năng điện toán hơn, chẳng hạn như bộ nhớ và sức mạnh xử lý. Điều này khiến các mô hình hồi quy logistic trở nên lý tưởng đối với những tổ chức đang bắt đầu với các dự án Machine Learning để đạt được một số thành tựu nhanh chóng.

**Sự linh hoạt**: Bạn có thể sử dụng hồi quy logistic để tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn. Bạn cũng có thể sử dụng phương pháp này để xử lý trước dữ liệu. Ví dụ: bạn có thể sắp xếp dữ liệu với một phạm vi giá trị lớn, chẳng hạn như giao dịch ngân hàng, thành một phạm vi giá trị hữu hạn, nhỏ hơn nhờ hồi quy logistic. Sau đó, bạn có thể xử lý tập dữ liệu nhỏ hơn này với các kỹ thuật Machine Learning khác để phân tích chính xác hơn.

**Khả năng hiển thị**: Phân tích hồi quy logistic cung cấp cho nhà phát triển khả năng nhìn nhận các quy trình phần mềm nội bộ rõ hơn so với các kỹ thuật phân tích dữ liệu khác. Khắc phục sự cố và sửa lỗi cũng trở nên dễ dàng hơn do các phép toán ít phức tạp hơn.

* Nhược điểm:

**Không thể xử lý được các tương tác non-linear giữa các đặc trưng**: Hồi quy logistic chỉ có thể xử lý được các tương tác tuyến tính giữa các đặc trưng. Nếu có tương tác non-linear giữa các đặc trưng thì kết quả phân loại của thuật toán có thể không chính xác.

**Dễ bị overfitting**: Khi số lượng đặc trưng quá lớn, hồi quy logistic dễ bị overfitting. Điều này đặc biệt xảy ra khi các đặc trưng có độ tương quan cao với nhau.

**Khó khăn trong việc xử lý dữ liệu thiếu**: Hồi quy logistic yêu cầu đầy đủ dữ liệu để huấn luyện, và nếu có sự thiếu sót trong các giá trị của dữ liệu, thuật toán có thể không chạy được hoặc cho kết quả sai lệch.

**Phụ thuộc vào đặc trưng đầu vào**: Hồi quy logistic phụ thuộc rất nhiều vào đặc trưng đầu vào và không thể hoạt động tốt nếu không có đặc trưng phù hợp. Việc chọn đặc trưng đầu vào cũng ảnh hưởng rất lớn đến kết quả của thuật toán.

### Ứng dụng của hồi quy Logistic

Hồi quy logistic có một số ứng dụng thực tế trong nhiều ngành công nghiệp khác nhau:

**Sản xuất:** Các công ty sản xuất áp dụng phân tích hồi quy logistic để ước tính xác suất xảy ra sự cố ở bộ phận trong máy móc. Sau đó, họ sẽ lên lịch bảo trì dựa trên xác suất đã ước tính này để giảm thiểu sự cố trong tương lai.

**Chăm sóc sức khỏe:** Các nhà nghiên cứu y khoa lên kế hoạch điều trị và chăm sóc dự phòng bằng cách dự đoán khả năng mắc bệnh ở bệnh nhân. Họ sử dụng các mô hình hồi quy logistic để so sánh tác động của tiền sử gia đình hoặc của bộ gen lên bệnh tật.

**Tài chính:** Các công ty tài chính phải phân tích các giao dịch tài chính để đề phòng gian lận, xem xét các đơn xin vay và đơn bảo hiểm để đề phòng rủi ro. Những vấn đề này phù hợp với mô hình hồi quy logistic bởi chúng có kết quả cụ thể, chẳng hạn như rủi ro cao hoặc rủi ro thấp và gian lận hoặc không gian lận.

**Bộ phận Tiếp thị:** Các công cụ quảng cáo trực tuyến sử dụng mô hình hồi quy logistic để dự đoán xem người dùng sẽ nhấp vào một quảng cáo hay không. Kết quả là, các nhà tiếp thị có thể phân tích phản ứng của người dùng đối với những từ ngữ và hình ảnh khác nhau, tạo ra các quảng cáo hiệu suất cao có khả năng thu hút khách hàng.

# Kết quả và đánh giá mô hình

## Phân tích chi tiết bài toán

### Tập dữ liệu

#### Nguồn dữ liệu

* Tên: Tập dữ liệu dự đoán bênh tiểu đường ở nữ giới ít nhất 21 tuổi của Ấn Độ Pima
* Nguồn: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

#### Giới thiệu

Tập data chứa các thông tin về bệnh tiểu đường ở nữ giới ít nhất 21 tuổi của Ấn Độ Pima

Tập data bao gồm 9 đặc tính (Feature):

* **Pregnancies**: mang thai .
* **Glucose**: lượng đường.
* **BloodPressure**: huyết áp.
* **SkinThickness**: độ dày da. Insulin: một hormone peptide được sản xuất bởi các tế bào beta của đảo tụy.
* **BMI**: chỉ số khối cơ thể.
* **Diabetes** Pedigree Function: chức năng phả hệ bệnh tiểu đường.
* **Age**: tuổi.
* **Outcome**: kết quả(1-bị tiểu đường/0-không bị tiểu đường).

Tập data gôm 768 trường dữ liệu

## Các công cụ, thư viện dùng xây dựng

Google BigQuery ML là một dịch vụ của Google BigQuery cho phép người dùng tạo và triển khai các mô hình học máy (machine learning models) trực tiếp từ dữ liệu lưu trữ trong BigQuery. Điều này giúp cho việc xây dựng và triển khai các mô hình học máy trở nên đơn giản và hiệu quả hơn.

Các tính năng chính của Google BigQuery ML bao gồm:

* Tích hợp dữ liệu: BigQuery ML cho phép người dùng sử dụng dữ liệu được lưu trữ trong BigQuery để xây dựng các mô hình học máy, tránh việc phải di chuyển dữ liệu sang một nơi khác để tiền xử lý hoặc huấn luyện mô hình.
* Hỗ trợ nhiều loại mô hình: BigQuery ML cung cấp hỗ trợ cho nhiều loại mô hình học máy như Linear Regression, Logistic Regression, K-Means clustering, và TensorFlow.
* Không cần kinh nghiệm về lập trình: Người dùng không cần có kinh nghiệm về lập trình để sử dụng BigQuery ML. Thay vào đó, họ có thể sử dụng các công cụ trực quan để xây dựng và triển khai các mô hình học máy.
* Tự động tối ưu: BigQuery ML có thể tự động tối ưu các siêu tham số của mô hình để đạt được hiệu suất tốt nhất.
* Thực hiện trên dữ liệu lớn: BigQuery ML có khả năng thực hiện huấn luyện và triển khai mô hình trên các tập dữ liệu lớn với hàng tỷ hoặc thậm chí hàng trăm tỷ dòng trong thời gian ngắn.
* Tích hợp với công cụ phân tích: BigQuery ML có thể tích hợp với các công cụ phân tích dữ liệu khác trên Google Cloud Platform như Google Data Studio để trực quan hoá kết quả của mô hình học máy.

Ở đây chúng ta dùng mô hình Hồi quy Logisstic (Logistic Regression), để xây dựng và dự đoán

## Xây dựng mô hình sử dụng BigQuery UI

### Đưa dữ liệu lên Google BigQuery

A screenshot of a computer

Description automatically generated with medium confidence

Hình . Chọn ADD

Chọn **ADD**

A screenshot of a computer

Description automatically generated with medium confidence

Hình . Chọn Local file

Chọn **Local file**

A screenshot of a computer

Description automatically generated with medium confidence

Hình . Điền thông tin table

Chọn file csv tải lên, project, dataset và table rồi **Create Table**

Sau khi tải lên thành công

A screenshot of a computer

Description automatically generated with medium confidence

Hình . Sau khi đưa dữ liệu lên thành công

### Xây dựng mô hình

Sử dụng CREATE MODEL của Google BigQuery ML để xây dựng mô hình. Ta dùng từ row 100 trở đi để xây dựng mô hình

CREATE OR REPLACE MODEL `diabetes.logistic\_reg\_model`

OPTIONS

  (model\_type='logistic\_reg',

    input\_label\_cols=['Outcome']) AS

SELECT

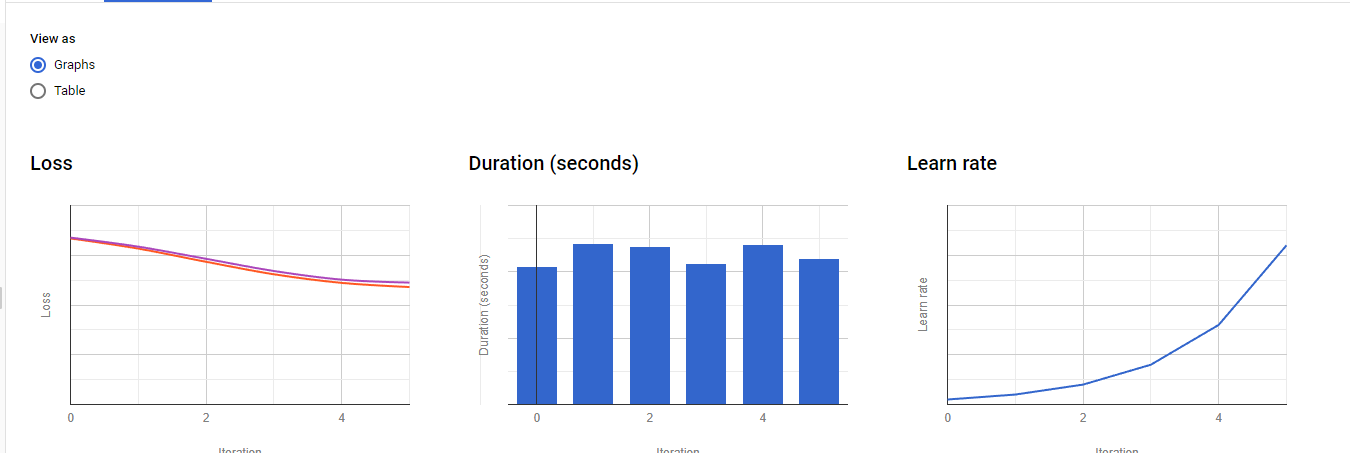
  \*

FROM

  `diabetes.data`

LIMIT 800 OFFSET 100

Kết quả sau khi thực thi SQL



Hình . Tổng quan về model

A screenshot of a chat

Description automatically generated with low confidence

Hình . Tổng quan về model



Hình . Tổng quan về model

Có thể thấy chỉ số chính xác của mô hình khoảng 76%, recall khoảng 52% và precission khoảng 82% đối với dữ liệu train

### Chạy thử mô hình

Sử dụng ML.PREDICT dự đoán mô hình. Ở đây ta dụng 100 dòng đầu của dữ liệu đã train đẻ làm dữ liệu test

SELECT

  predicted\_Outcome, raw.Outcome as real\_OutCome

FROM

  ML.PREDICT(MODEL `graduationprojectms-75dd9.diabetes.logistic\_reg\_model`,

    (

    SELECT

      \*

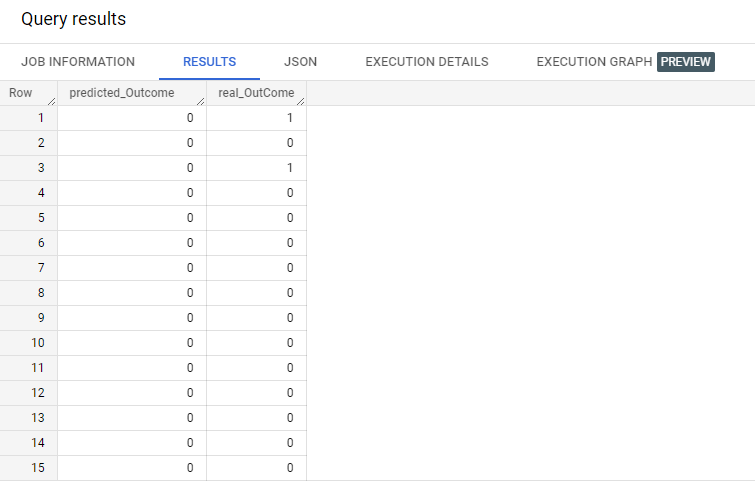
    FROM

      `diabetes.data`)),

      `diabetes.data` as raw

LIMIT 100

Kết quả thực thi



Hình . Kết quả thực thi

### Đánh giá độ chính xác của mô hình

Sử dụng ML.EVALUATE để đánh giá so với dữ liệu. Ở đây ta dụng 100 dòng đầu của dữ liệu đã train đẻ làm dữ liệu test.

SELECT

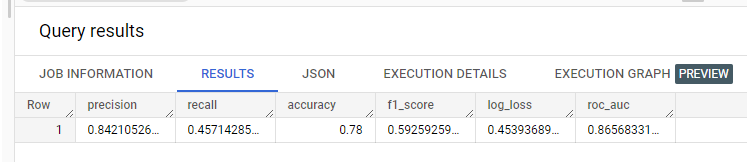
  \*

FROM

  ML.EVALUATE(MODEL `graduationprojectms75dd9.diabetes.logistic\_reg\_model`,

    (SELECT \* FROM`diabetes.data` LIMIT 100))

Kết quả thực thi



Hình . Kết quả thực thi

Ta có thể thấy độ chính xác dự đoán là 78%, recall khoảng 45% và precission khoảng 84% đối với dữ liệu test

### Ma trận nhầm lẫn

Sử dụng ML.CONFUSION\_MATRIX để tạo ma trận. Ở đây ta dụng 100 dòng đầu của dữ liệu đã train đẻ làm dữ liệu test.

SELECT

  \*

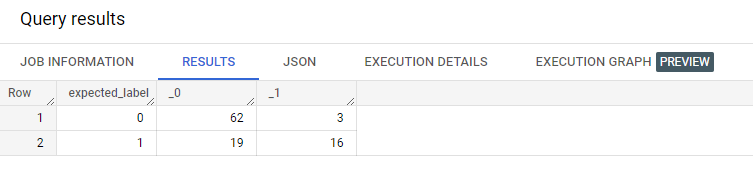
FROM

  ML.CONFUSION\_MATRIX

(MODEL `graduationprojectms75dd9.diabetes.logistic\_reg\_model`,

    (SELECT \* FROM`diabetes.data` LIMIT 100))

Kết quả thực thi



Hình . Kết quả thực thi

Kết Luận

Từ việc xây dựng và đánh giá model được trình bày trong chương 3, chúng ta có thể đưa ra một số kết luận như sau:

- Độ chính xác (Accuracy Score) của model là 76% cho tập train và 78% cho tập test.

- Độ chính xác này có thể chấp nhận được bởi lớn hớn 70% và mức độ chênh lệch độ chính xác giữa tập train và tập test là không nhiều (2%) do vậy tỉ lệ bị overfit của model được giảm thiểu.

- Tỷ lệ recall khoảng 52% cho mô hình train và 47% cho mô hình test có nghĩa là tỷ lệ dự đoán mắc bệnh đúng là 50%

- Tuy nhiên, một phần lý do dẫn đến độ chính xác không quá cao là bởi tập dữ liệu khá nhỏ (chỉ có 768 mẫu) nên khi chia tập train-test theo tỉ lệ 85-15, model có thể sẽ không train được tất cả các trường hợp đặc biệt.

- Ngoài ra, để cải thiện độ chính xác của model, chúng ta có thể sử dụng một số thuật toán khác trong Machine Learning, ví dụ như Non-Linear Regression , …

TÀI LIỆU THAM KHẢO

[Machine Learning cơ bản (machinelearningcoban.com)](https://machinelearningcoban.com/2017/08/31/evaluation/)

[BigQuery documentation  |  Google Cloud](https://cloud.google.com/bigquery/docs)

[Diabetes Dataset | Kaggle](https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset)

[3.1. Hồi qui Logistic — Deep AI KhanhBlog (phamdinhkhanh.github.io)](https://phamdinhkhanh.github.io/deepai-book/ch_ml/classification.html)

<https://kyanon.digital/google-bigquery-la-gi-tim-hieu-ve-google-bigquery/>

<https://aws.amazon.com/vi/what-is/logistic-regression/>

<https://cloud.google.com/bigquery/?utm_source=google&utm_medium=cpc&utm_campaign=japac-VN-all-vi-dr-BKWS-all-super-trial-PHR-dr-1605216&utm_content=text-ad-none-none-DEV_c-CRE_602335001443-ADGP_Hybrid%20%7C%20BKWS%20-%20PHR%20%7C%20Txt%20~%20Data%20Analytics%20~%20BigQuery_BigQuery-bigquery-KWID_43700071566407023-aud-1596662388894%3Akwd-301529154162&userloc_9040331-network_g&utm_term=KW_bigquery&gclid=Cj0KCQjwr82iBhCuARIsAO0EAZzqzstKTQMoehDLhuEyhpqTbGDW_h_lm5hFsFjnmpeZrRUjjUsuKJ4aAtvaEALw_wcB&gclsrc=aw.ds>

<https://devsamurai.vn/vi/blog-vi/tong-quan-ve-google-bigquery/>

<https://luanvan1080.com/hoi-quy-logistic.html>

PHỤ LỤC